Exercise: Understanding Language Models

Goals The main goal of this worksheet is for you to understand language models a bit more before you implement them.

Question 1 Consider the prefix "*LeBron James talked about* _". Think about the different n-gram orders here: $n = 2$ (1 word of context for the language model, just *about*), through $n = 5$ (all four words of context). **Do you think $n = 3$ or $n = 4$ will yield the same distribution over next words as $n = 5$? Why or why not?**

Question 2 Consider the following corpus (collection of sentences), extended from the one in the video:

I like to eat cake but I want to eat pizza right now. Mary told her brother to eat pizza too. He went to Pizza Hut to get some.

What is the probability distribution of words following *to* under a 2-gram model? That is, what is $P(y \mid \text{to})$? Hint: this should be a list of words, each one associated with a probability. You don't need to explicitly write down all of the words with zero probability.

Question 3 What data structure or data structures would you use to store the words and probabilities for $P(y \mid \text{to})$?

(Optional) Question 4 Now suppose you were going to store the entire 2-gram model: the words and probabilities $P(y \mid x)$ for every (x, y) pair. What data structure or data structures would you use for this?